PROCEEDINGS OF SCIENCE

# Data scouting and data parking with the CMS high level trigger

**Swagata Mukherjee**[*][†]
*III. Physikalisches Institut A, RWTH Aachen University, Germany*
*E-mail:* s.mukherjee@cern.ch

The CMS experiment has devised two new strategies at the high level trigger to search for new physics in difficult corners of the phase space, or in large samples with B hadrons. The first strategy, called Data Scouting, introduced in Run 1, allows to take data that would otherwise be rejected by the normal trigger filters. It is based on event-size reduction rather than event filtering and it is useful, for instance, to search for low mass resonances. The second strategy, called Data Parking, aims at overcoming the main limitation in the CMS data taking, which is the computing power involved in the prompt reconstruction. In 2018, a large amount of additional data, more than $1 \times 10^{10}$ events containing a pair of B hadrons, was collected by CMS and parked for a delayed offline reconstruction during the Long Shutdown 2. This dataset was triggered requiring a soft displaced muon originating from the decay of a B hadron, without applying any selection on the other B hadron, allowing an unbiased sample for competitive measurements on rare B-meson decays. Both methods are reviewed here.

*European Physical Society Conference on High Energy Physics - EPS-HEP2019 -*
*10-17 July, 2019*
*Ghent, Belgium*

[*] Speaker.
[†] On behalf of the CMS Collaboration.

PoS(EPS-HEP2019)139

## 1. What is data scouting?

Data scouting is a paradigm for LHC data analysis based on trigger-level event reconstruction. It complements the traditional analysis paradigm in which events are selected by trigger system and sent for an expensive prompt offline reconstruction procedure. By taking advantage of the online reconstruction that already takes place at the high level trigger (HLT), it is possible to significantly increase the number of physics events stored for analysis while having negligible impact on computing resources.

## 2. Why do we need data scouting?

There are several constraints on the number of events that can be recorded using standard high level triggering framework in CMS. (1)The data acquisition system (DAQ) of CMS has finite bandwidth; so restrictions on the data volume are imposed by the size of the temporary raw data storage at LHC Point 5 (the site of the CMS experiment), and by the bandwidth of the link between Point 5 and the CMS computing center at the main CERN site. (2)The prompt reconstruction system must be able to reconstruct all selected events promptly without significant backlog. It is desired that all physics data be reconstructed and available within 48 hours of being collected. (3)The total amount of storage space (tape and disk) for data is limited. The cost of purchasing storage space needs to be considered when deciding how much data to record. (4)The trigger decision at the HLT must be made within a few hundred milliseconds.

As a consequence of these constraints, CMS records events for physics analysis at an average rate of approximately 1 kHz, which is several orders of magnitude smaller than the rate (up to 40 MHz) at which pp collisions occur in the detector. It would be meaningless to record every single collision event, because most do not contain interesting physics and would not be used by any analysis. Nevertheless, the requirements of the trigger impose significant constraints on current searches and measurements. These constraints become more aggressive over time as LHC luminosity rises.

For example, many searches for new physics in hadronic final states rely on $H_T$ triggers. The $H_T$ variable is a proxy for the amount of hadronic activity in the event and is often used as a new physics search variable. In the beginning of Run II, the loosest $H_T$ trigger in the HLT menu selected events having $H_T > 800$ or 900 GeV. Events with $H_T$ below the threshold cannot be recorded unless they present some other feature of interest.

The HLT performs reconstruction algorithms similar to those used offline. This includes a version of the PF algorithm and its components: track finding, clustering of calorimeter energy deposits, and muon, electron, photon, and hadron identification. The good performance of the physics objects produced by these algorithms suggests a new strategy for analyzing CMS data:

- Events are reconstructed at real time in HLT by running the PF algorithm or other reconstruction algorithms. Apply a loose selection on the reconstructed physics objects. For each event passing the loose selection, save the HLT-reconstructed physics objects to disk.
- Discard the raw data.
- Perform searches for new physics using the saved HLT-level events.

This strategy is referred to as data scouting [7]. While the full raw data for a CMS event is around 1 MB in size, the physics objects reconstructed by the HLT can be represented using only a few kB of memory. Trigger-reconstructed events can therefore be recorded at rates of several kHz and still not occupy more DAQ bandwidth than a single ordinary HLT path. Since the objects reconstructed at trigger level is saved in the dataset and to be used in analyses, this entirely removes the need for prompt reconstruction for the selected events. This means that the limit on event processing time from the offline reconstruction system is not relevant. Since the raw data is discarded for scouting, storage space is only needed for the reduced dataset consisting of the reconstructed trigger objects. This requires a negligible amount of disk resources compared to storing full events.

## 3. $H_T$ scouting

The idea of data scouting was conceived in Run I of CMS, and the technique was used to perform searches for exotic resonances decaying to dijets. Dijet resonance searches tend to be severely constrained by trigger requirements: events with two back-to-back jets are very common at hadron colliders and a trigger that recorded all of them would have extremely high rate. In CMS, dijet searches traditionally use $H_T$ triggers. The $H_T$ trigger threshold rises with increasing LHC energy and luminosity, and this implies that the lowest resonance mass the search can probe is pushed higher and higher over time. Data scouting provides dijet searches with relief from the rising $H_T$ trigger rates. This was first demonstrated in 2011, when a data scouting trigger path was deployed in CMS for the last few hours of data taking at 7 TeV [1].

After the successful demonstration, a second scouting trigger was designed and deployed at the HLT for most of the 2012 CMS data taking period. The trigger required $H_T > 250$ GeV and had a rate of around 1kHz. This rate was too high for the PF algorithm to be run for every event, so the trigger instead reconstructed and saved calorimeter jets (calo jets), which are clustered directly from energy deposits in the ECAL and HCAL. Calo jets require negligible HLT resources to reconstruct, and at high momentum their mass resolution is adequate despite the lack of tracking information. A dijet search was conducted using the calo jet scouting dataset of 2012, and strong limits were obtained for dijet resonance with intermediate mass [2]. A small fraction of the scouting data is also saved in the standard event format, to allow a jet-by-jet comparison of the offline and online reconstruction performances. The energies of online and offline reconstructed jets agree within around 2% in the kinematic phase-space of the dijet search analysis.

For Run II of CMS it was desired to build a more comprehensive software framework for data scouting that would enable a variety of physics analyses. The following strategy was adopted in 2015, in the beginning of Run II.

- Run the HLT PF algorithm and record PF jets and particle candidates at the maximum attainable rate
- Beyond that, reconstruct and record calo jets instead

Because the event content of the PF algorithm is different from (and much larger than) that of calo jet reconstruction, two scouting data formats were designed, one for each flavor of reconstruction. The data formats are denoted the calo-scouting and PF-scouting event formats, and they were designed to be as lightweight as possible.

The typical calo-scouting event format had an average size of approximately 1.5 kB, while the PF-scouting event size was around 10 kB, most of which is occupied by the PF candidate objects. The inclusion of the PF candidates allows for more complex analysis strategies involving, for example, jet substructure variables computed using the constituents of the jets. The PF jet scouting trigger was successfully used in a search for pair-produced three-jet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV [5], and the calo jet scouting trigger was once again used for dijet search in 13 TeV [3, 4].

## 4. Muon scouting

The PF-scouting stream also contained a trigger that selects events having two muons, with no requirement on jet activity in the event. The muons are required to have $p_T > 3$ GeV and to have dimuon invariant mass $m_{\mu\mu} > 10$ GeV. This dimuon scouting trigger was implemented at the beginning of Run II as a proof-of-concept of a non-hadronic scouting trigger.

The success of the $H_T$ scouting framework in 2015 and 2016 prompted interest from data analysts seeking to expand the range of new physics searches possible with CMS data. One possibility that was pursued was to use data scouting to collect events with two muon candidates and to perform a search for dark photons decaying to muons.

To build a scouting trigger for a dark photon search, the dimuon trigger availble in 2015 and 2016 scouting trigger menu had to be modified significantly. The changes were intended to loosen the kinematic selection on the muons, enabling muon pairs with low $p_T$ , low invariant mass, and possible displacement from the beamline to be selected. The major changes were:

- The L1 requirement was loosened substantially. The 2016 muon scouting trigger's L1 seed imposed moderate $p_T$ thresholds on both leading and subleading muons. It was supplemented in 2017 with a large number of new L1 seeds, including some with minimal $p_T$ requirements; these instead require the muons to be close together in $\eta$ and to lie within the CMS barrel. Some L1 seeds place an opposite-charge requirement on the muon pair.
- The dimuon invariant mass requirement was removed.
- Quality requirements on the dimuon vertex to be close to primary vertex were removed, enabling the selection of muon pairs from displaced vertex.

The muon objects used in the scouting data format were modified to include more detailed information about the muon's associated track, namely the values and uncertainties of the parameters defining the track, and muon quality information. These additional variables are needed in the context of the targeted dark photon search to select well-reconstructed muon pairs. Additionally, scouting vertex objects were updated with a more complete set of uncertainty information. Secondary vertices from displaced muon decays were reconstructed by the trigger path and saved in the scouting event content. Finally, to handle the increased rate resulting from loosening the L1 requirement and removing the dimuon mass cut, the modified trigger path was moved from the PF-scouting stream into the calo-scouting stream. The PF reconstruction sequence was removed from the trigger path, and the calo-scouting stream was reconfigured to save HLT muon objects.

The dimuon invariant mass spectrum for events selected by the new trigger in 2017 and 2018 is shown in Figure 1.

Because of the removal of the mass cut, the dimuon invariant mass spectrum extends far below 1 GeV. The known resonances in the $m_{\mu\mu}$ spectrum can be seen clearly. A small excess of events around 330 MeV corresponds to $\phi \to K^+K^-$ decays where Kaons are misidentified as prompt muons.

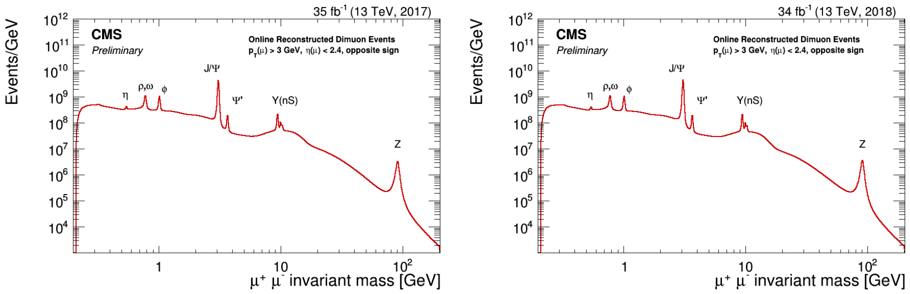

**Figure 1:** Dimuon invariant mass spectrum reconstructed in the High Level Trigger system of the CMS detector for 2017 data (full) and 2018 data (partial).

A search for a narrow resonance decaying to a pair of muons is performed in 11.5-45.0 GeV mass range using muon scouting data [6]. The results of this search are interpreted in the context of a dark photon ($Z_D$) that may feebly couple the standard model (SM) particles to a hidden, dark sector of particles. The dark photon interacts with SM particles through the kinetic mixing of its $U(1)_D$ gauge field with the $U(1)_Y$ hypercharge field of the SM. The degree of this mixing, and consequently, the strength of the coupling of $Z_D$ with SM fermions is determined by the kinetic mixing coefficient $\varepsilon$. The search sets strong constraints on dark photon mass and mixing. It is possible to perform searches for even lower mass resonances and long-lived new particles giving rise to displaced dimuon signature, using the data collected by the dimuon scouting trigger.

## 5. Data parking

The term data parking refers to the technique of selecting events at the HLT and immediately moving them to tape storage, skipping prompt reconstruction. Events selected in this way remain on tape until there are sufficient free computing resources to reconstruct them. Data parking allows more than the standard 1 kHz of physics events to be recorded, because the rate is not constrained by the limited capacity of the prompt reconstruction system. The achievable rate of data parking is constrained by the bandwidth of the CMS DAQ and by the amount of tape storage space. Data parking fits naturally into the data scouting paradigm, especially in the context of searches for new physics. A main drawback of data scouting is the discarding of the raw data and the reliance entirely on HLT-reconstructed physics objects, which may suffer more from detector noise and miscalibration than the standard offline-reconstructed objects. If a search for new physics is performed on scouting data and if a new signal is found, it may be difficult to find out whether the supposed new physics is real or the result of noise that affected the HLT reconstruction in some way. Parking the raw events selected by the data scouting triggers removes this problem. If a physics result with the scouting data is called into question, the parked data can be brought from tape to disk and reconstructed. The analysis can then be performed again using the parked data to confirm or disconfirm the result. In case the scouting analysis returns an unambiguous result, the parked data can stay on tape forever, sparing the resources that would be needed to reconstruct it. This strategy was deployed in CMS in 2015 and 2016 to complement the data scouting triggers. It was estimated that the DAQ could handle several hundred Hz of parked data safely. A suite of parking triggers was created, along with a new data stream to hold them. These triggers select the same events as the triggers in the PF-scouting stream, dividing them among a number of parked datasets. In 2017, the rate allocated to scouting triggers had increased significantly, mostly due to the new dimuon trigger, and it was not possible to park all scouting data. So the parking triggers were replaced by prescaled versions of the scouting triggers, which recorded only 10% of scouting events.

In 2018, the parking strategy had significantly changed. CMS collaboration is interested in performing lepton universality tests by measuring $R_{K^*}$, which is the ratio of the branching fractions of $B^0 \to K^* \mu^+ \mu^-$ and $B^0 \to K^* e^+ e^-$. While CMS already had triggers in place to select a sufficiently large sample of $B^0 \to K^* \mu^+ \mu^-$, there were no triggers that would allow to collect an adequate $B^0 \to K^* e^+ e^-$ data sample. So, in 2018, full data parking capability was utilised to collect such non-muon final states. More than 10 billion events containing a pair of B hadrons were recorded during 2018, and this big dataset was triggered by requiring a soft and displaced muon originating from the decay of a B hadron, without applying any selection on the other B hadron, allowing an unbiased sample for competitive measurements of rare B-meson decays. The parked data is now being reconstructed and will be analysed soon.

## References

[1] CMS Collaboration, “Search for narrow resonances using the dijet mass spectrum in pp collisions at $\sqrt{s} = 7$ TeV,” CMS-PAS-EXO-11-094, https://cds.cern.ch/record/1461223.

[2] V. Khachatryan *et al.* [CMS Collaboration], “Search for narrow resonances in dijet final states at $\sqrt{(s)} = 8$ TeV with the novel CMS technique of data scouting,” Phys. Rev. Lett. **117**, no. 3, 031802 (2016) doi:10.1103/PhysRevLett.117.031802 [arXiv:1604.08907 [hep-ex]].

[3] A. M. Sirunyan *et al.* [CMS Collaboration], “Search for dijet resonances in proton–proton collisions at $\sqrt{s}$ = 13 TeV and constraints on dark matter and other models,” Phys. Lett. B **769**, 520 (2017) Erratum: [Phys. Lett. B **772**, 882 (2017)] doi:10.1016/j.physletb.2017.09.029, 10.1016/j.physletb.2017.02.012 [arXiv:1611.03568 [hep-ex]].

[4] A. M. Sirunyan *et al.* [CMS Collaboration], “Search for narrow and broad dijet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on dark matter mediators and other new particles,” JHEP **1808**, 130 (2018) doi:10.1007/JHEP08(2018)130 [arXiv:1806.00843 [hep-ex]].

[5] A. M. Sirunyan *et al.* [CMS Collaboration], “Search for pair-produced three-jet resonances in proton-proton collisions at $\sqrt{s}$ =13 TeV,” Phys. Rev. D **99**, no. 1, 012010 (2019) doi:10.1103/PhysRevD.99.012010 [arXiv:1810.10092 [hep-ex]].

[6] CMS Collaboration, “Search for a narrow resonance decaying to a pair of muons in proton-proton collisions at 13 TeV,” CMS-PAS-EXO-19-018, https://cds.cern.ch/record/2684861.

[7] D. Anderson [CMS Collaboration], “Data Scouting in CMS,” PoS ICHEP **2016**, 190 (2016) doi:10.22323/1.282.0190.